Executive Summary: Problematic Internet Usage

Aaron Weinberg, Emilie Wiesner, Dan Visscher
GitHub: https://github.com/aarondweinberg/CMI_problematic_internet_use/

Introduction

Internet use has been identified by researchers as having the potential to rise to the level of addiction, with associated increased rates of anxiety and depression. Identifying cases of problematic internet usage currently requires evaluation by an expert, however, which is a significant impediment to screening children and adolescents across society. One potential solution is to rely on data that is more easily and uniformly collected: the kind collected by a family physician, by a simple survey, or by a smartwatch. The research question this project sets out to answer is: "Can we predict the level of problematic internet usage exhibited by children and adolescents based on their physical activity and survey responses?"

Dataset

The data comes from a research study conducted by the Child-Mind Institute, with data for 3960 participants. The target variable is a severity impairment index (SII) that measures problematic internet use on a scale from 0 (no impairment) to 3 (severe impairment). There are 33 predictor variables, including demographics (e.g., age, sex), physical measurements often taken by a family physician (e.g., height, weight, blood pressure), results of a fitness test (e.g., sit & reach, endurance time), survey responses and scales (e.g., internet usage in hours per day, sleep disturbance, children's global assessment), and measures from a bio-electric impedance analysis (e.g., bone mineral content, fat mass index). Additionally, about one month of accelerometer data was provided for almost one thousand of the participants in 5-second intervals. Significantly, participants often completed some parts of the study but not others, so any given participant will have entire groups of variables missing.

Preprocessing

Participants were dropped if they did not have an SII score (about a third of the participants). The distribution of remaining SII scores is very skewed: over half of participants had an SII score of 0, while only 30 (about 1% of) participants who were measured had an SII score of 3. For each input variable, we found benchmarks to identify extreme values and removed these. Because many variables had high correlation with other variables, we engaged in several rounds of feature reduction and identified important features using a random forest.

Model Selection and Results

We compared models using Cohen's kappa function. This function measures the accuracy of prediction for ordinal variables, with random guessing producing a score of 0, scores of 0.4-0.6 indicating moderate or fair agreement levels, and scores above 0.75 producing excellent agreement.

We designed an iterative imputer for missing predictor values, a function transformer to compute activity zones for various predictors, and an algorithm to sequentially apply classifiers to ordinal data. We then ran a pipe that first imputed and computed predictors, oversampled the sii=3 class using a Synthetic Minority Oversampling TEchnique, and made predictions using a variety of models. Cohen's kappa indicated that using gradient boost to predict PCIAT scores and then modified bins to convert to sii scores had the best kappa score of 0.448.

For our final model we used a **tuned gradient boosting regressor**. This took our test data and used it to predict PCIAT values and then used custom-tuned bins to compute SII values. This resulted in a **kappa value of 0.456**, suggesting that our model avoided overfitting but was only able to come up with a moderate amount of agreement.